Malicious Insider Threat Detection: A Conceptual Model

Tesleem Fagade and Theo Tryfonas

{tesleem.fagade, theo.tryfonas}@bristol.ac.uk

Cryptography Group, Faculty of Engineering
University of Bristol
Bristol, United Kingdom

Abstract

The advent of Internet technologies, growing number of sophisticated hacking tools and mobile workforce creates a new dimension to the malicious insider problem for many organisations. In spite of the significant interest from researchers and industry experts, trusted employees with elevated access continue to pose insider challenges to organisation risk mitigation efforts. It is suggested that malicious insiders show certain personality traits, leave behind digital footprints and observable cyber risk behaviour in advance of an attack. This work offers a different perspective to address the insider problem by drawing concepts from behavioural theory, personality profiling and digital trails auditing. Instead of isolated treatments, our approach considers the intersection of different risk domains and aggregates risk scores from each as a predictor of malicious insider activities. This model has significant implication for security professionals, to draw insight from inextricably linked risk domains within the context of cybersecurity management. However, substantial empirical work is still needed to evaluate the model in real world cases.

Keywords: malicious insider, insider threat detection, cybersecurity, information security, cyber risk behaviour.

1 Introduction

Modern economies, the dynamics of how organisations create, process and distribute information; and the increased reliance on large-scale interconnected information systems also lead to increased cyber-risk exposure [1]. Information security is, therefore, a critical issue for organisations. In this cyber battle, an agent's elevated access to information and services becomes potentially a vulnerability to the system, signifying the issue of the insider threat. As suggested in the relevant literature, insiders are the weakest link in organisations' defence posture [2][3]; a system is more susceptible to insider threats exploitations compared to exploitations due to failures of technical and procedural measures. In addition, it is difficult to estimate insider problem because many organisations fail to report insider misuse [4]. Hence, a signi-

ficant challenge is the design and development of an automated and cost effective insider threat preventive tool, which can distinguish malicious activities in advance of an attack.

The advent of internet technology and sophisticated hacking tools and mobile workforce creates a thriving environment for malevolent employees and a whole new dimension to the malicious insiders' problem. Insiders misuse of information systems mostly centres around destroying, modifying and stealing corporate data [4], and most of these illicit acts are carried out through technical capabilities. A study [5] of 36 illicit cyber activities in the government sector suggests that 24% of incidents is due to unauthorised privilege users, of which 11% involves installation of backdoors. Similarly, in a case study of 52 cyber incidents [6], it is shown that 57% are detected through system irregularities of which 73% involves remote access logs and 57% involves unauthorised file access logs. There are limitations to the effectiveness of technical controls in preventing motivated insiders from carrying out malicious activities in an organisation [7], which is why academic researchers continue to investigate individual differences in personality traits, as a precursor for malevolent behaviour within the context of information security.

Research linking cybersecurity compliance and human factor has expanded over the years, in an attempt to establish how individual differences shapes security policy compliance intentions. Security-related behaviour in organisation continues to generate research interest in the information security literature. Behavioural theories provide guidelines on how behaviour may manifest at different stages of an insider threat scenario by recognising observable 'concerning behaviour' that insider exhibits in advance of security exploits [8]. The theory of planned behaviour suggests that a person's intention, perceived behaviour towards crime, subjective norms and attitude are key factors in predicting behaviour [9]. A study [10] shows that in 23 cases of insider threat in the banking and finance sector, 33% is due to personal problems that are unrelated to employment, like breakup and anxiety; 23% is due to revenge, 27% is due to debt and 81% is due to financial gains. Also, based on a case study of 52 illicit cyber activities in the IT and Telecommunication sector, report [6] suggests that 33% is due to intolerant to criticism, 57% involves disgruntled employees, 47% is revealed through overt behaviour, and 58% involves direct communication of threat online.

The high complexity of information systems along with the socio-technical nature of insider threats require a robust mechanism that can fuse numerous detector alerts for the analysis of patterns as a precursor to threats. This is particularly important because when activities are performed to varying degrees by both benign and malicious insiders, the accuracy of isolated detector alert becomes uncertain. Therefore a holistic interdisciplinary approach that blends technological, psychological and organisa-

tional elements of the insider threat problem is required [11]. This work is aimed at developing a conceptual model of a malicious insider detection system that can combine threat indicators from different inter-related domains in order to prevent cyber risk while lowering false-positive thresholds and background noise. We based each risk indicator on a hypothesis about malicious insider personality traits, behavioural patterns and technical footprints, supported with evidence from academic literature and subject experts' opinion. We draw concepts from the theory of planned behaviour, the 'big five' personality dimensions and technical anomaly detection, such that, different elements of threat that are inextricably linked from different domains is a function of the malicious insider problem. This model could be useful in ordering highly significant developing insider threats that require security analysts' review and could also provide further insights for collective management of information security in organisations.

The structure of this report is as follows; section 2 covers related work and the motivation for this work. The background description of our conceptual model is presented in section 3. Method and preliminary design are covered in section 4, while section 5 covers the conclusion and future work.

2 Related Work

While it is not a new area of research, the work in the domain of insider threat prediction has continued to generate interest among researchers. Previous studies indicate the inadequacy of attempts to address the human factor in cyber security despite the evidence that a malicious insider exhibits an observable 'concerning behaviour' in advance of the actual exploit [8]. There have been a variety of approaches to solving the malicious insider problem, some studies are based on behavioural analysis of a malicious insider to understand the key drivers of malevolent intentions. For instance, [12] combines real-time technical data information obtained from users' information systems with their psychometric test score data, then applies decision algorithm in order to assess their predisposition to malicious behaviour. In a related study [13], an analytical model to simulate adversary behaviour is presented, it describes malicious insider threats through a devised taxonomy of attributes like access, risk, knowledge, process and motivation, in order to analyse how each or a combination of these attributes stimulate malicious insiders' behaviour. Also [14], applies cognitive dissonance and neutralisation theory to insider perception and behaviour, to establish a link between rationalisation and the likelihood of an insider to commit a crime. Some literature also emphasises the importance of recognising early signs of risky behaviour. In that light, a predictive model that can assist with risk mitigation decision is presented by [15], the work describes an automated model for the support and detection of high-risk behavioural indicators. Similarly, the work

proposed by [16], evaluates the probability of IT misuse through the lens of multi-level hierarchical layers of threat components, the work provides an insight on how users' behaviour at the system level can be harnessed to predict computer misuse that originates from legitimate users.

In terms of technical assessment, [17] shows how technical tools like the intrusion detection/prevention systems (IDS/IPS), and log analysis can be leveraged to uncover insider activities, while [18] describes how insider threat problems can be mitigated through resilience, survivability and security by combining the technical assessment of information asset categorization with agents behaviour categorization. In terms of personality attributes, some research emphasises the link between personality traits and the tendency to become a malicious insider. For instance, [19] suggests that the personality trait of narcissism is a common characteristic of malicious insiders, while [7] presents a personality evaluation approach that links cyber-security protocol violation to the personality trait of a malicious insider, depending on deterrence, protection motivation or efficacy factors. While recognising that insider threat constitutes a problem which already has damaging consequences for organisations, insider threats cannot always be detected or appropriately addresses with technical solutions alone [20], and there is a need for a framework that encompasses multiple risk indicators for a holistic and predictive threat detection [21][22]. This work recognises the important contribution of other models, however, to the best of our knowledge, we believe that our model offers a different perspective to detecting malevolent activities by considering the intersection of personality traits, behavioural risks and technical footprints indicators as a single block. This analytical model will allow security managers to draw insights from inextricably linked variables from independent domains, to address insider problems within the context of cybersecurity management.

3 Background Description of the Conceptual Model

An important consideration here is that due to the introductory and analytic modelling of this work, this paper did not provide in-depth analysis of the quantifying metrics and weighting factors assigned to each insider threat indicators. It should be noted that assigned weights are subjective and depend on the expertise of the human resources, data analytics, cybersecurity teams and the risk tolerance of each organisation. In table 1, weighting factors associated with risk factors from different domains are provided, such tables are maintained by the administrator from each domain. In the case of technical and behavioural risk indicators, weighting values are assigned based on the alert status/count associated with each risk element; while scenario effect parameters are the determinants of weighting values in the case of personality risk factor. For instance, in the technical risk domain,

unlicensed/unauthorised software installation requires just a single count (one occurrence) and a weighting factor of 0.38. The rest of this section provides a brief description of each domain and risk factors considered for this work.

Personality	**Notation**	**Security Scenario Effect**	**Weight**
Openness	O	Low sense of sanction severity	0.35
Conscientiousness	C	Low sense of response efficacy	0.21
Extraversion	E	Low sense of threat severity, threat vulnerability and response cost	0.45
Agreeableness	A	Low sense of sanction certainty	0.25
Narcissism	N	Low sense of sanction certainty	0.65
Technical	**Notation**	**Alert Status/Count**	**Weight**
Unauthorized logins	UL	1	0.28
VPN/Remote logins	VRL	2	0.36
Unauthorised Software Installation	USI	1	0.38
File Deletion/Modification	FDM	1	0.32
Viruses/Malware	VM	3	0.25
Behaviour	**Notation**	**Alert Status/Count**	**Weight**
Destructive Behaviour	DB	2	0.38
Intolerance to Criticism	IC	3	0.43
Security Violation	SV	2	0.32
Drug and Alcohol Abuse	DAA	3	0.45
Isolation and Seclusion	IS	4	0.28

Table 1: Risk indicators.

3.1 Personality Risk Factors

There are different methods and assessment tools like the Myers-Briggs Type Indicator (MBTI), NEO-PI-R, survey questions and social media profiling that can be utilised to measure personality traits based on the big-five psychological construct of Openness, Conscientiousness, Extroversion, Agreeableness and Neuroticism (OCEAN). Then each element of OCEAN can be related to a significant aspect of behaviour. Although, personality trait is fairly stable throughout an individual's lifetime and it could be a key determinant of intentions but studies suggest that

individuals with the same personality react differently to the same condition depending on associated security scenarios like self-efficacy, sanction severity, sanction certainty and response cost [23]. The personality risk factors aspect of this work is mainly based on the empirical studies conducted in [7], [23] and we summarise this in table 1 by showing the cross-level relationship that exists between personality types and security scenario effects. Weight is assigned to the personality/security scenario interaction based on behavioural intention to violate cybersecurity protocol.

3.2 Technical Risk Factors

There are varieties of security tools [24][25][26]that can be deployed for network and host level assessment. Although, most of these tools are based on audit process that performs security checks against known vulnerabilities but they can provide a snapshot of technical risk factors within an organisation. For instance, tools like the SIEM/log analysis and Intrusion Detection Systems (IDS) can provide sufficient information on the changes to configuration file binaries, access authentication logs and other anomalies from which a logical connection can be drawn about a user's activities [27]. This can allow system administrators to profile a normal user based on job roles and privileges, such that, if there is an irregular pattern in the log information for a particular user, compared to a normal user for the same role, then that may be an indication of a potential malevolent insider activities. Table 1 shows some of the hypothesised cyber risks included in this model, indicating the alert status (count) for each situational risk factor and the corresponding weighting factor.

3.3 Behavioural Risk Factors

Though some behavioural risks like disgruntles, destructive behaviour and policy violations could be triggered by external factors that are unrelated to employment or criminal backgrounds like anxiety, breakup, depression and debt but may help address psychological factors required to form group homogeneity [19], [28]. Behaviour and external environmental influences can indicate early signs of cybersecurity risks, and the more an individual exhibits one or more combinations of the behavioural risk elements, the more likely it is to violate cyber security protocols. Human resource staff are particularly well trained to apply observation techniques, recognise and report high scoring risk indicators as a predictor of anomalous behaviour. Behavioural risk indicators, alerts status and associated weight factor considered for this study, is shown in table 3.

4 Method and Preliminary Design

The model considers different types of risk situational factors in an organisation i.e. behavioural, personality and technical/system level risk elements from three different domains; the human resource (H.R), personality profiling and technical (I.T) domain respectively. In line with evidence from literature, we hypothesise that when the technical risk factor is used in isolation, it is susceptible to false positives and cannot be a strong indication of malicious insider activities. However, when an agent is suspected of violating organisation security protocols, other situational risk factors can be aggregated to detect with increased accuracy, the pattern of activities that can characterise a malevolent insider. Using this approach, organisations can build employee profile and define a 'normal' risk threshold (R.T), based on threat indicators and the job role for that employee. To determine R.T, consider the high-level abstractions for the conceptual model shown in figure 1. Data flow from the 3 domain streams are combined and processed for a single value output. Data from social network platforms, surveys, 360-degree profiler and other personality tests can be leveraged to obtain personality trait for an employee. Also, by constantly monitoring and analysing behavioural risks and psychological state for that employee, the Human resource (H.R) can provide data for a given time, indicating the behavioural risks for that employee. Similarly, incident log data obtained from IT department can provide input for the employee technical risk indicators. These inputs can then be analysed to determine an acceptable R.T for that employee.

Employee activities are then monitored over a period e.g. monthly and risk score (R.S) is compared to a defined risk threshold (R.T). If there is a significant deviation from the 'normal' pattern in each time period, the system flags warning triggers risk-status for that employee and the employee is placed under close supervision; thereby invoking the organisation's standard threat mitigating procedure. If risk threshold is not met or exceeded, risk-status is not triggered and observation simply continues.

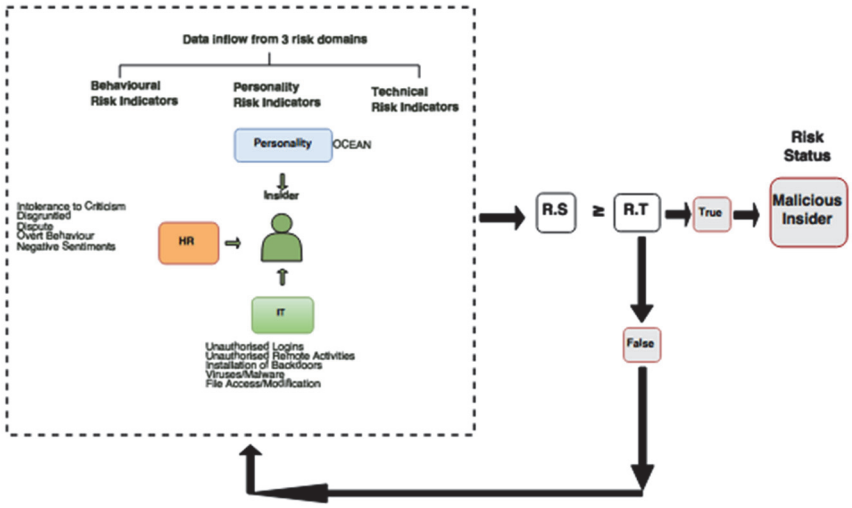

Figure 1: High-level abstraction of the insider threat modelling.

The key assumption in this model is how weighting values are assigned to risk elements of the threat indicators, as shown in table 2. For instance, we combine the personality traits from the 'OCEAN' construct and scenario level association for each of the 'OCEAN' risk elements, to obtain cross-level interactions for the intention to commit computer abuse, as described in [23]. Then a weighted value is assigned to each personality trait based security risk associated with each personality type, as described in the previous section 3. This step is then repeated for behavioural and technical risk indicators, such that, risk elements (R_e) from each risk domain have corresponding weights (R_{eW}), as shown in table 2.

Domain		Risk Factors				
$p_{risk\text{-}i}$	R_e	**O**	**C**	**E**	**A**	**N**
	R_{eW}	0.35	0.21	0.45	0.25	0.65
$b_{risk\text{-}i}$	R_e	**DB**	**IC**	**SV**	**DAA**	**IS**
	R_{eW}	0.38	0.43	0.32	0.45	0.28
$t_{risk\text{-}i}$	R_e	**UL**	**VRL**	**USI**	**FDM**	**VM**
	R_{eW}	0.28	0.36	0.38	0.32	0.25

Table 2: Risk Elements.

To determine the security risk status for an employee and decide if such an employee constitute insider risk to an organisation, consider the high-level algorithm for the malicious insider threat detection in table 3. Firstly, organisations decide the value of R.T for an employee, which is the number corresponding to acceptable risk baseline for that employee. Each of the risk domains, i.e. personality risk indicators ($p_{risk\text{-}i}$), technical risk indicators ($t_{risk\text{-}i}$) and behavioural risk indicators ($b_{risk\text{-}i}$) have associated R_e. The input to the model has combined R_e from each of the $p_{risk\text{-}i}$, $b_{risk\text{-}i}$ and $t_{risk\text{-}i}$ domains. However, each R_e has associated and predetermined weighting factor R_{eW}. Also, each employee has a state for a given period with respect to R_e, such that; if the state of R_e is TRUE, the corresponding value for R_{eW} is returned. The output from the model is a numeric value *R.S*. For a given period, *R.S* is obtained by aggregating all the R_{eW} for that employee. If *R.S* is greater or equal to the *R.T* set for that employee, then it is assumed that insider threat activity is detected. Then the employee is flagged for further investigation, otherwise, normal monitoring simply continues.

Algorithm

Personality risk indicators = $p_{risk\text{-}i}$
Behavioural risk indicators = $b_{risk\text{-}i}$
Technical risk indicators = $t_{risk\text{-}i}$
Risk-threshold (R.T) = acceptable baseline allowed for an employee.
Risk Score (R.S)
Risk Elements = R_e
Risk Element Weights = R_{eW}

STEP 1: FOR each EMPLOYEE:
There is an input ($p_{risk\text{-}i}$, $b_{risk\text{-}i}$, $t_{risk\text{-}i}$);
There is an output (R.S);

To determine R.S for an EMPLOYEE:
Organisation decide values for R.T: 1.55
Make a list of all possible R_e = {r_1, r_2, r_3……..r_n}
Make a list of corresponding R_{eW} = {w_1, w_2, w_3……w_n}

There is a STATE = TRUE, FALSE (1, 0)
Such that, in each month:

STEP 2: FOR each employee risk assessment ($p_{risk\text{-}i}$, $b_{risk\text{-}i}$, $t_{risk\text{-}i}$)
IF the State for R_e = 0
Return (NULL)
IF the State for R_e = 1
Return (R_{eW});
R.S = ∑(R_{eW});

IF R.S $\geq$ *R.T*
Insider threat is detected;
Flag employee as malicious;
ELSE IF R.S $\neq$ *R.T*
Return to STEP 1;

Table 3: High-Level Algorithm for the Malicious Insider Threat Detection

The preliminary simulation is done in MATLAB by using matrix tables for all true states of R_e and corresponding R_{eW}. Simulation output of our model is shown in figure 2. As independent variable changes over the 12-month period, the user risk profile is shown as the sum of the R_{eW} from the three risk indicators associated with that user, in each time period. For illustrative purposes, the *R.T* for the employee under consideration is set at 1.55. In month 7, the user risk profile is above the *R.T* value of 1.55, set for that user; indicating that insider threat is detected and the user is flagged as malicious. In month 7, personality trait 'O' is recorded for the user based a chosen profiling method. HR records 3 or more counts of 'DAA', 2 or more counts of 'DB' and 4 or more counts of 'IS'. Similarly, IT security record shows 2 or more counts of 'VRL'. By aggregating all the risk indicators and comparing *R.S* to the user's *R.T*, is shows early warnings of malicious insider activities.

Crucially though, treating each of the threat indicators in isolation is not enough to trigger the user as malicious, even though there may be false positive indicators in places. Consider month 11, HR $b_{risk\text{-}i}$ record for the same user has a risk indicator value of 1.11, quite a negligible difference to the record for month 7, which has a $b_{risk\text{-}i}$ value of 1.09. However, the IT $t_{risk\text{-}i}$ recorded in month 11 has a risk indicator value of zero, which is significantly different to the record for month 7, which has a risk indicator value of 0.36. The $p_{risk\text{-}i}$ for month 11 and month 7 is also similar, at risk indicator values of 0.25 and 0.35 respectively. Therefore, if the user is assessed based on $b_{risk\text{-}i}$ alone, activities for month 7 and 11 would be flagged, which could trigger false positive alert. However, combining all threat indicators from different domains shows a significant difference in the user's activities for both months, and only the activity of month 7 exceeds R.T.

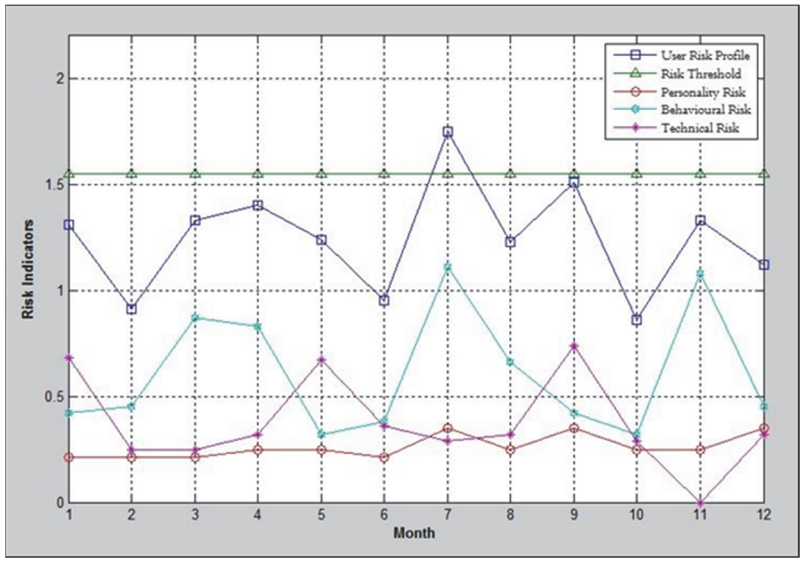

Figure 2: Simulation output showing how to detect malicious insider activities from multiple risk indicators

5 Conclusion

The starting point for managing organisation risk is the insider threat assessment, upon which management can build extra layers of controls like policy guidelines, awareness training and technical security solutions. We propose an approach that can get valuable insight into the activities of malicious insider so that organisations can have more information to draw inferences about insider actions during an investigation. This work describes an analytical model that considers risk elements from different risk domains, such that when each domain is treated in isolation, it leads to insufficient evidence of malicious intention. However, when the intersection of different risk indicators is considered as a single block, it offers a considerable improvement to the possibility of detecting an insider threat. The results in this paper have been explained with reference to theories and past literature and we suggest that future research examines the relationship and key attributes that specifically link evidence of malicious intentions from technical log information, human behaviour and personality traits. As part of future work, we hope to empirically explore this model and evaluate its sustainability in real world cases like banking organisations.

References

[1] J. Hua and S. Bapna, "Who Can We Trust?: The Economic Impact of Insider Threats," *J. Glob. Inf. Technol. Manag.*, vol. 16, no. 4, pp. 47–67, 2013.

[2] S. Aurigemma and R. Panko, "A Composite Framework for Behavioral Compliance with Information Security Policies," in *2012 45th Hawaii International Conference on System Sciences*, 2012, pp. 3248–3257.

[3] T. Fagade and T. Tryfonas, "Security by Compliance? A Study of Insider Threat Implications for Nigerian Banks," Springer, Cham, 2016, pp. 128–139.

[4] R. Chinchani, D. Ha, A. Iyer, H. Q. Ngo, and S. Upadhyaya, "Insider Threat Assessment: Model, Analysis and Tool," in *Network Security*, Boston, MA: Springer US, 2010, pp. 143–174.

[5] Kowalski Dawn et al., "Insider Threat Study: Illicit Cyber Activity in the Government Sector," 2008.

[6] E. Kowalski and D. Cappelli, "Insider Threat Study: Illicit Cyber Activity in the Information Technology and Telecommunications Sector," 2008.

[7] M. Mcbride, L. Carter, and M. Warkentin, "Exploring the Role of Individual Employee Characteristics and Personality on Employee Compliance with Cybersecurity Policies," 2012.

[8] F. L. Greitzer, R. E. Hohimer, and A. Biography, "Modeling Human Behavior to Anticipate Insider Attacks," *J. Strateg. Secur.*, vol. 4, no. 2, pp. 25–48, 2011.

[9] NCCIC, "Combating the Insider Threat," *Natl. Cybersecurity Commun. Integr. Cent.*, no. May, pp. 61–64, 2014.

[10] A. Cummings, T. Lewellen, D. Mcintire, A. P. Moore, and R. Trzeciak, "Insider Threat Study: Illicit Cyber Activity Involving Fraud in the U.S. Financial Services Sector," 2012.

[11] D. D. Caputo, M. A. Maloof, and G. D. Stephens, "Detecting insider theft of trade secrets," *IEEE Secur. Priv.*, vol. 7, no. 6, pp. 14–21, 2009.

[12] M. Kandias, A. Mylonas, N. Virvilis, M. Theoharidou, and D. Gritzalis, "An Insider Threat Prediction Model," in *TrustBus - International Conference on Trust, Privacy and Security in Digital Business*, 2010, pp. 26–37.

[13] B. Wood, "An insider threat model for adversary simulation," *SRI Int. Res. Mitigating Insid. Threat to Inf. Syst.*, vol. 2, pp. 1–3, 2000.

[14] K. Padayachee, "An Insider Threat Neutralisation Mitigation Model Predicated On Cognitive Dissonance (ITNMCD)," *South African Comput. J.*, vol. 56, no. 56, pp. 50–79, Jul. 2014.

[15] F. L. Greitzer and R. E. Hohimer, "Modeling Human Behavior to Anticipate Insider Attacks," *J. Strateg. Secur.*, vol. 4, no. 2, pp. 25–48, 2011.

[16] G. B. Magklaras and S. M. Furnell, "Insider threat prediction tool: Evaluating the probability of IT misuse," *Comput. Secur.*, vol. 21, no. 1, pp. 62–73, 1998.

[17] S. Zeadally, B. Yu, D. H. Jeong, and L. Liang, "Detecting Insider Threats: Solutions and Trends," *Inf. Secur. J. A Glob. Perspect.*, vol. 21, no. 4, pp. 183–192, 2012.

[18] K. Roy Sarkar, "Assessing insider threats to information security using technical, behavioural and organisational measures," *Inf. Secur. Tech. Rep.*, vol. 15, no. 3, pp. 112–133, 2010.

[19] M. Kandias, K. Galbogini, L. Mitrou, and D. Gritzalis, "Insiders trapped in the mirror reveal themselves in social media," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2013, vol. 7873 LNCS, pp. 220–235.

[20] S. Steele and C. Wargo, "An Introduction to Insider Threat Management," *Inf. Syst. Secur.*, vol. 16, no. 1, pp. 23–33, Mar. 2007.

[21] E. E. Schultz, "A framework for understanding and predicting insider attacks," *Comput. Secur.*, vol. 21, no. 6, pp. 526–531, 2002.

[22] F. L. Greitzer, D. A. Frincke, and M. Zabriskie, "Social/Ethical Issues in Predictive Insider Threat Monitoring."

[23] M. Warkentin, M. McBride, L. Carter, A. Johnston, and A. C. Johnston, "The Role of Individual Characteristics on Insider Abuse Intentions," *Assoc. Inf. Syst. AIS Electron. Libr.*, no. 1, 2012.

[24] OpenVAS, "OpenVAS - Open Vulnerability Assessment System: The world's most advanced Open Source vulnerability scanner and manager," *openvas.org*, 2016. [Online]. Available: http://www.openvas.org/. [Accessed: 12-Mar-2017].

[25] Tenable Network Security, "Nessus Vulnerability Scanner," 2017. [Online]. Available: https://www.tenable.com/products/nessus-vulnerability-scanner. [Accessed: 12-Mar-2017].